

# **Bridging the Gap: Enhancing Prominence and Provenance of NASA Datasets in Research Publications**

Irina Gerasimov<sup>a,b</sup>, Andrey Savtchenko<sup>a,b</sup>, Jerome Alfred<sup>a,b</sup>, James Acker<sup>a,b</sup>, Jennifer Wei<sup>b</sup>, Binita KC<sup>a,b</sup>

<sup>a</sup>ADNET Systems Inc.

<sup>b</sup>NASA GES DISC

Corresponding Author [Irina.Gerasimov@nasa.gov](mailto:Irina.Gerasimov@nasa.gov)

## *Address:*

Goddard Earth Sciences Data and Information Systems

NASA Goddard Space Flight Center

Code 619

Greenbelt, MD 20771 USA

## *Authorship Statement:*

Irina Gerasimov - developed the idea, contributed to tools and methodology development, and drafted the manuscript.

Andrey Savtchenko, Jerome Alfred, James Acker, and Jennifer Wei - contributed to tools and methodology development and revised the manuscript.

Binita KC - revised the manuscript.

## **Abstract**

Attribution of datasets that were used to generate research results described in peer-reviewed publications to the original source of these datasets (which are often archived at NASA Earth Science data centers) has been very challenging. Even though the data citation standard of citing datasets as research artifacts and citing them with Digital Object Identifiers (DOIs) was introduced over a decade ago, most authors do not properly reference the data used in their studies and merely mention them in the text. The lack of proper citations of datasets makes the peer-reviewed publication less transparent, imperils reproducibility, and impedes open science. We offer an open-source *publication management methodology and a tool* that can help to enhance usage-based data discovery, prominence, and provenance of the data; reproducibility of the research results; and potentially increase the return on investment on NASA-funded research.

**Keywords:** open-source, data provenance, data usage, data publication, dataset citation, dataset DOI, citation management

## **Introduction**

Both government agencies and academic institutions produce a vast volume of data that is openly accessible and free for public use. Thousands of scientific journal research articles have been published by analyzing these datasets. It is paramount to track how the datasets have been used within the research community to understand and verify the return on investments in these datasets produced by numerous publicly-funded research programs. Although the practice of journals requiring persistent and resolvable references to data has recently been improving, the non-citation of datasets remains a persistent and sizable challenge. As a result, existing bibliographic databases, such as Scopus and Web of Science, don't have sufficient direct linkages built between the datasets used in an article with dataset DOIs, which uniquely indicate the original source of the data. Linking datasets with the publication increases the reproducibility and transparency of the research outcomes and enhances Findability, Accessibility, Interoperability, and Reusability (FAIR) principles in data management and stewardship (Wilkinson et al., 2016). Conducting analyses using Earth Science data often requires expert users who have an advanced understanding of scientific data formats and algorithms, as well as knowledge of the fundamental elements of Earth Science. However, this poses several challenges for data centers with diversified user communities, especially those from various geographic regions, and those possessing different levels of education, expertise, and research background. The authors' institution has created a library of peer-reviewed publications that directly links data collections and their research applications. Making relevant publications a prominent aspect of the dataset documentation can guide non-expert users of the data in discovering and narrowing down the data and variables to meet their specific needs. This collaboration has developed an open-sourced method and tool that can automatically extract peer-reviewed publication records from bibliographic sources to create a publication library with the data-to-publication linkage.

## Current State of Linking Datasets and Documents

The emergence of dataset citation in the late 1990s has been influenced by the emphasis on open science, data sharing, and reproducibility, giving credit to dataset creators and providing metrics for dataset use (Costas, 2013; Parsons, 2019; Robinson-Garcia, 2016; Wilkinson, 2016). The first principles of data citations were established by CODATA-ICSTI (2013), indicating that dataset citations should have *the same significance as other citations in the scholarly record*. To be considered article-quality, datasets must have metadata, undergo peer review, be searchable and discoverable in databases, and be cited (Costas, 2013; Kratz, 2015). Despite these efforts, links between datasets and documents still need to be improved due to the low citation of datasets in the research literature (Mooney, 2012; Robinson-Garcia, 2016; Park, 2017; Silvello, 2018, Zhao, 2018, Vannan, 2020). Reports have shown that over 85% of datasets covered by Web of Science's Data Citation Index were uncited (Robinson-Garcia, 2016; Peters, 2016).

Dataset DOI registries such as DataCite (Brase, 2009) provide persistent identifiers for research data, making it easier to discover and cite datasets. They also provide metadata to help users understand the context and content of the datasets. The dataset registry metadata is used by a number of services to discover and link datasets, including the Web of Science's Data Citation Index.

Initiative for Open Abstracts (I4OA) and Initiative for Open Citations (I4OC) are initiatives led by DataCite to promote data sharing and open science. I4OA advocates for open access to scholarly literature, while I4OC encourages publishers to make their citation metadata openly available. Another initiative promoting open science and data sharing is Scholix (Cousijn, 2019) which aims to link research data and literature by creating a framework for exchanging information between scholarly communication platforms. These initiatives contribute to the growing movement towards open science and data sharing, which is crucial for advancing scientific research and innovation.

Dataset repositories and data-sharing initiatives are important tools for promoting open data and making research more transparent and reproducible. By enabling researchers to easily find and reuse

datasets, these initiatives help to increase the impact and visibility of research and facilitate collaboration between researchers in different fields.

In addition to Web of Science's Data Citation Index, other major services linking datasets and documents are Scopus (Burnham, 2006), Crossref, and Google Scholar. Web of Science and Scopus are popular subscription-based bibliographic databases that cover a range of journals, books, and proceedings selected based on journal impact factor (Web of Science), or by independent subject boards or advisory experts (Scopus). Crossref is a non-profit organization that provides a Digital Object Identifier (DOI) registration service for scholarly content (Hendricks, 2020). Crossref provides infrastructure for creating and registering DOIs for various research outputs, such as journal articles, books, book chapters, conference proceedings, reports, preprints, and datasets. Crossref also operates the Crossref OpenCitations Index (COCI), which is a database of open DOI-to-DOI citations that allows for the tracking of citation links between research publications (Heibi, 2019). The dataset registry DataCite became a bibliometric database when it started providing the DOIs of the documents citing the datasets (Robinson-Garcia, 2017). DataCite dataset-to-document linkage is based on Crossref's COCI connections.

In contrast to the bibliographic databases described above, Google Scholar indexes documents rather than their sources (Van Norden, 2014; Prins, 2016). It has an undisclosed number of records, likely due to its constant improvement in indexing and searching (Gusenbauer, 2019). In addition to its comprehensive coverage, Google Scholar includes dissertations, books, and conference proceedings that are not covered by Web of Science or Scopus. However, there are cases where Google Scholar indexes non-scholarly content, such as book reviews, low-impact documents, non-refereed sources, and duplicate records that cause extra citations (Delgado López-Cózar, 2018; Halevi, 2017; Prins, 2016; Martin-Martin, 2018). The limited metadata provided by Google Scholar makes it difficult to use for bibliometric research (Martin-Martin, 2018; Halevi, 2017; Chapman, 2019).

Recently emerging OpenAlex (Priem, 2022) is a free, open-source scientific knowledge graph. It contains metadata for hundreds of millions of scholarly entities, including works, authors, and venues. OpenAlex provides a free API that allows search over article titles, abstracts, and texts, making it a

Google Scholar competitor. As of August 2023, the number of journal articles and books covered by OpenAlex is estimated to be 209 million, as opposed to Google Scholar, which is around 389 million. The OpenAlex API delivers exhaustive bibliographic details for the items retrieved, in contrast to Google Scholar which only provides URLs that require further processing to extract necessary bibliographic metadata.

The above-described bibliometric sources can be searched by the dataset DOI to find datasets linked to research documents. The dataset repository DataCite can also identify publications linked to datasets via DOIs of datasets registered with DataCite. These approaches require that datasets are cited with their DOI. Due to low dataset citation rates, automated approaches have been suggested to identify datasets using their mentions. Lane et al. (2020) summarized machine learning algorithms for detecting socioeconomic sciences data collections. Duan et al. (2018) proposed machine learning methods to identify data collection entities in Earth Science papers based on the names of instruments, missions, and major dataset variables. These approaches require access to the document texts, which can be complicated due to publisher paywalls, copyright restrictions, limited journal availability, language barriers, and publisher embargoes on articles. In addition, these automated approaches are still too immature to be the primary method of mapping data collections to research citations used in data center libraries, due to the ambiguity of referencing utilized datasets.

There are several citation management systems (CMS) currently available (Böhner and Teichert, 2020). Among those, the most widely used are the commercial products EndNote and RefWorks, and freely available products Mendeley (Li and Thelwall, 2012) and Zotero (Roy Rosenberg Center for History and New Media). The Zotero server, [zotero.org](https://www.zotero.org), is hosted on the Amazon Web Services (AWS) cloud. While Zotero server software is open source, the code only supports server-to-client interaction with the zotero.org server instance. There are several studies comparing these CMSs (Ivey and Crum, 2018). In the creation of a citation library at the Goddard Earth Sciences Data and Information Services Center (GES DISC), many factors were considered in the selection of the CMS, including platform support, cost, the versatility of citation export and import, content sharing between team members,

citation attachment (PDF) management, and the availability of an application programming interface (API), as well as versatility of citation tagging needed for linking citations and datasets.

To address the lack of dataset citations and obtain an exhaustive collection of document citations linked to NASA Earth Science datasets, we built a library of document citations based on Zotero CMS. The datasets used were all archived at the GES DISC, one of NASA's Earth Observing System Data and Information System (EOSDIS) data archive centers (Behnke, 2019). To demonstrate that a search of Google Scholar can provide counts of document citations that exceed searches of nominal bibliometric databases, we used over 1,500 GES DISC public datasets. These datasets have been registered with DataCite following NASA's Earth Science Data and Information System (ESDIS) Project DOI registration process (Wachoo, 2017) and have a common DOI prefix, 10.5067.

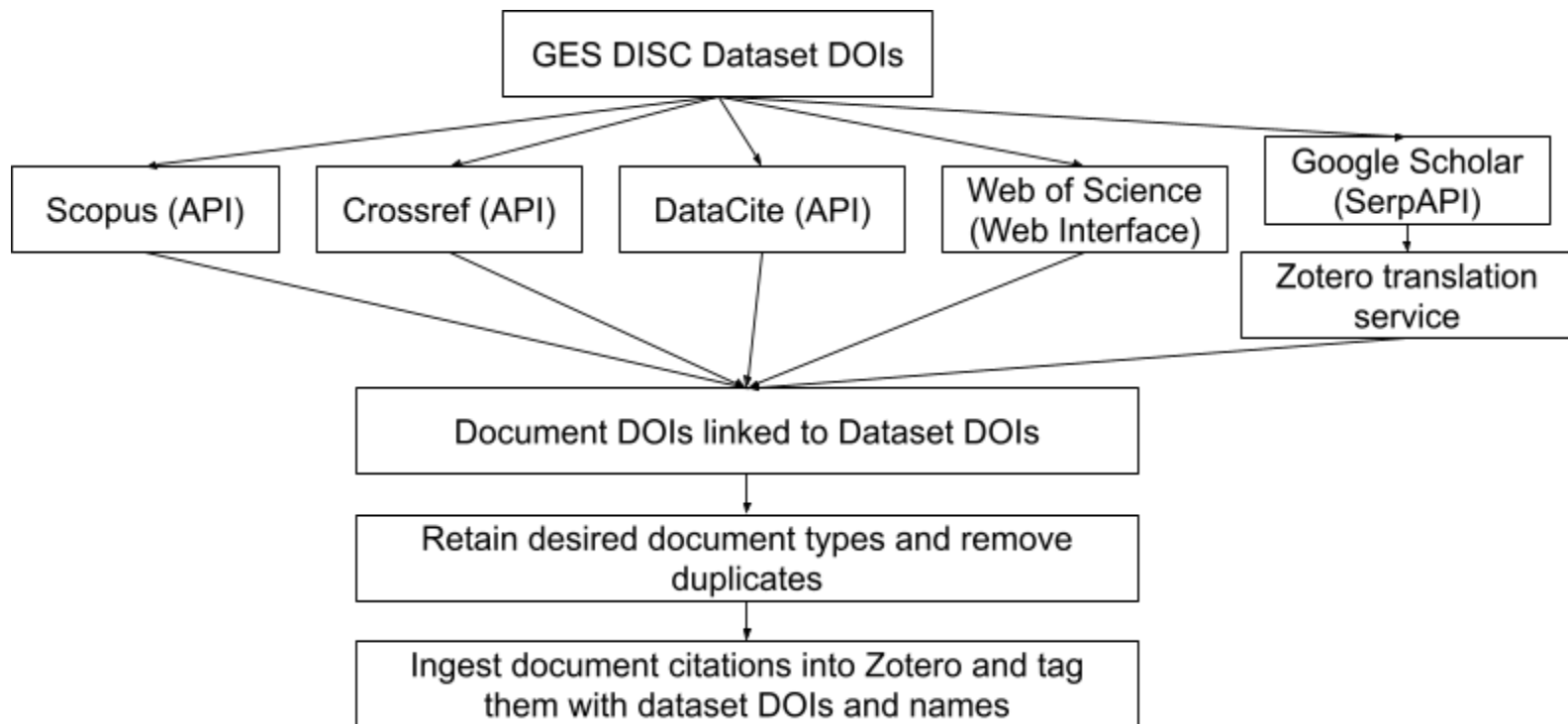
## **Study Contributions**

The study proposes an automated, high-precision solution for creating a trusted dataset-document citation linkage. This solution relies on a controlled dictionary that manages metadata for datasets and is used to search over Google Scholar. The proposed methodology for processing Google Scholar results allows for retaining deduplicated citations of specific document types such as books, book chapters, theses, journal papers, and conference proceedings. Collected document citations complement the dataset-document linkage which is created by searching Google Scholar and major bibliometric databases for dataset DOIs. Collected documents are added to the dataset landing pages and made available to the dataset users.

In addition, the study proposes a document-to-dataset linking approach that links various versions of datasets to all documents that reference any of those versions (and no versions in many cases). This approach ensures that research associated with prior dataset versions is not lost and is directly available to data users.

## Automated Citation Collection by Dataset DOI search

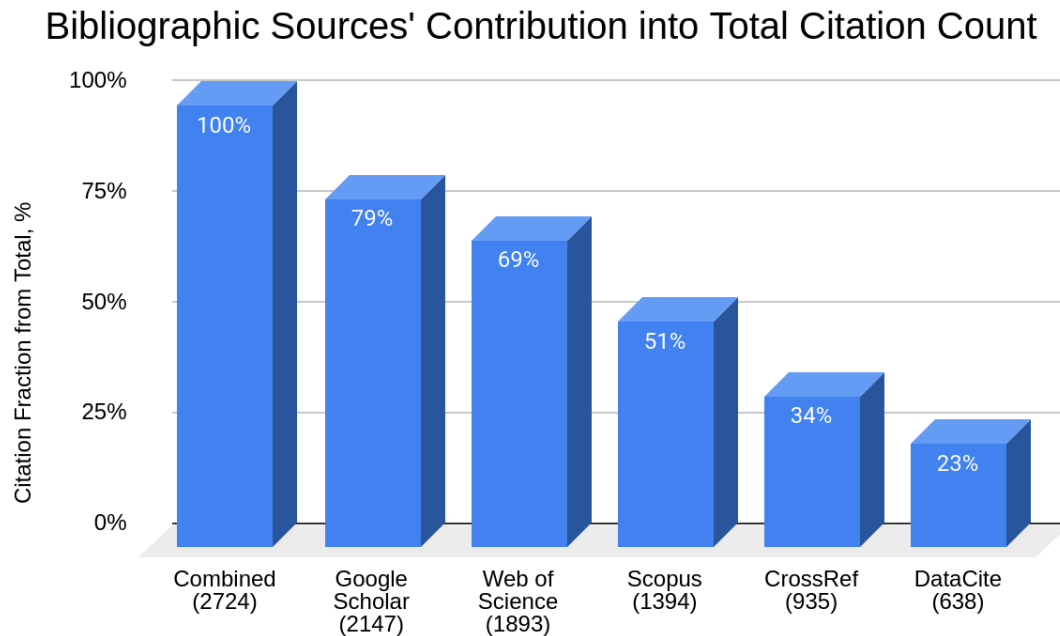
We utilized available Application Programming Interfaces (APIs) from Crossref, Scopus, and DataCite to automate the retrieval of documents citing the datasets. These APIs provide a list of document DOIs linked to the dataset DOIs. As the Web of Science API is not available, we used the Web of Science web interface to acquire citations and their references for a given dataset DOI. To improve efficiency, we searched the Web interface for the dataset DOI prefix, *10.5067*, and then processed resulting citations and their references to obtain dataset–document linkages. We used the subscription API, SerpAPI, from Google Scholar, which returns URLs of documents linked to the dataset DOIs. To determine the citations of these documents, we employed the Zotero translation service. To streamline the result merging process, only documents with DOIs were retained. We queried Crossref to determine the types of documents, which include books, book chapters, research articles, conference proceedings, theses, and reports. We excluded duplicate content by filtering out discussion papers and preprints. Figure 1 provides a diagram describing the dataset–document citation-acquiring process.





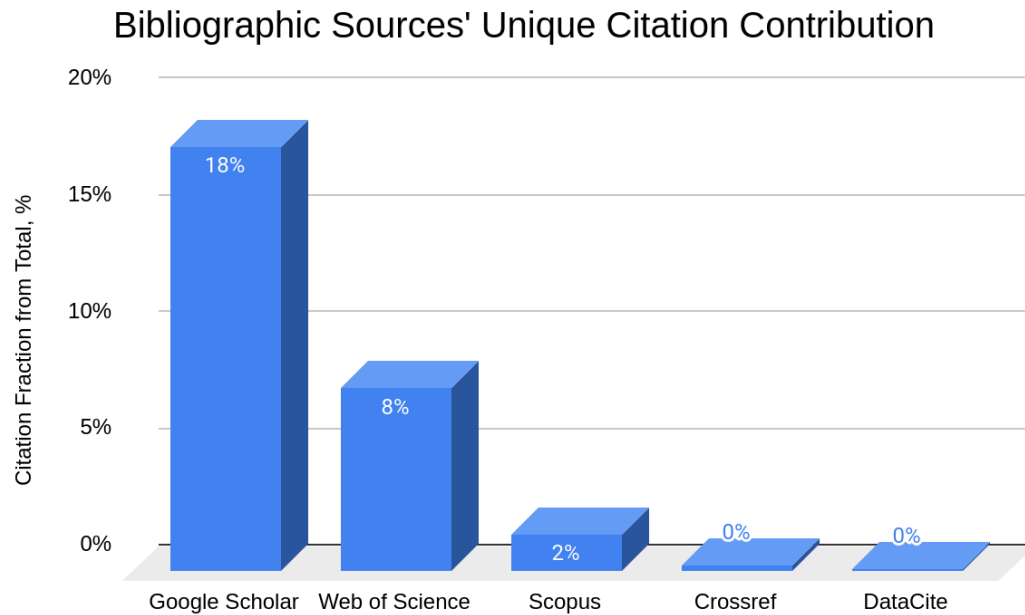
**Figure 1.** Flow diagram of obtaining dataset–document linkages from bibliographic sources

As of the middle of 2023, the total count of document citations found by all considered bibliographic sources is 2,724. Figure 2 shows a fraction of citations found by each individual source from the total citation count. As seen in Figure 2, 79% of the citations found by all considered sources are returned by Google Scholar.



**Figure 2.** Citations found in each bibliographic source as a percentage of total citations

Figure 3 shows the unique citation fraction contributed to the total citation count by an individual bibliographic source. As seen in Figure 3, Google Scholar contributes 18% of citations that are not provided by any other sources.



**Figure 3.** Percentage of unique citations per bibliographic source

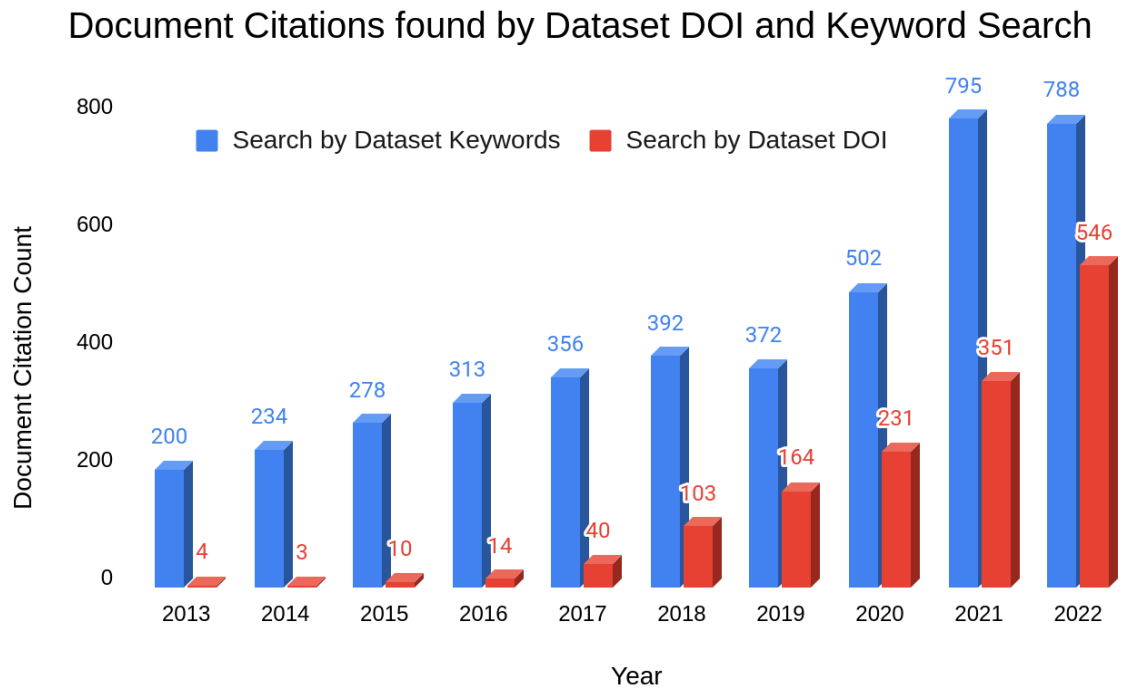
Our results demonstrate that out of all considered bibliographic sources, Google Scholar can return the largest number of citations with the largest count of citations not found in any other sources. This can be explained by several factors. Web of Science and Scopus have limited coverage due to their selectiveness. Crossref only recently started gaining access to citation references. DataCite relies on Crossref for the citation index, so it is even more limited than Crossref. Google Scholar indexes vast amounts of citations including international and less selective journals that might not be covered by Scopus and Web of Science.

### **Automated Citation Collection from Google Scholar**

In our work, we use Google Scholar for keyword searches because of its vast citation coverage, availability, and low cost of its API. With an increasing amount of open-access publications, Google Scholar gains access to those document texts. So instead of using difficult-to-access texts from publishers such as Elsevier and Springer, we use Google Scholar.

The objective of the keyword search is to automatically select keywords for the datasets so that these keywords return citations that are linked to these datasets with high certainty. While the datasets can be described by many keywords, the selection of the keywords should be done to obtain as many citations as possible, with minimal erroneous results due to dataset name ambiguity. The keywords should thus be specific enough to find publications linked to the dataset, and not to the same word that has the dataset name but which is not related to that dataset. For example, if we search for the Ozone Monitoring Instrument (OMI) bromine oxide (BrO) dataset “OMBRO” (Chance, 2007) we will get a lot of publications in the Portuguese or Spanish language (where “ombro” means “shoulder”) that did not use this dataset.

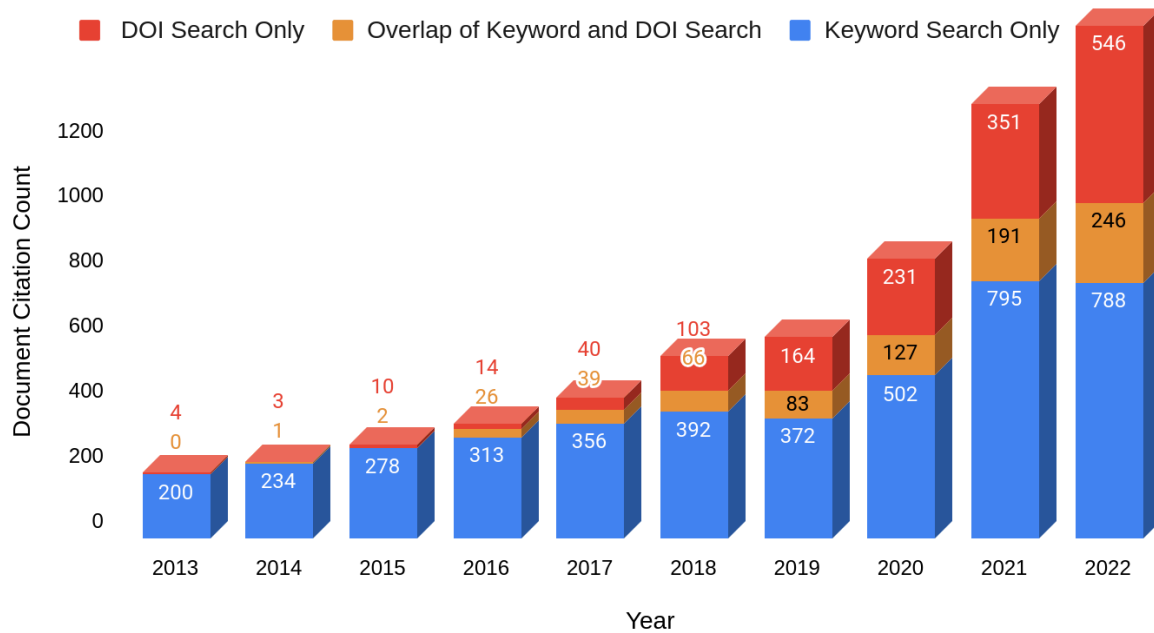
The GES DISC dataset collection metadata are stored in NASA's Earth Observing System Data and Information System (EOSDIS) Common Metadata Repository (CMR), which is governed by the Global Change Master Directory (GCMD) Keywords ontology. GCMD Keywords is a hierarchical set of controlled Earth Science vocabularies that allows for a consistent description of Earth Science datasets. The GCMD Keywords are organized into twelve sets, each describing Earth science keywords, platforms, instruments, data centers, locations, projects, services, and data resolution. Using these keywords and the dataset's unique name can help narrow down document citations. For the Google Scholar query, we use the dataset's essential metadata, such as the dataset's unique name defined by Earth Science Data Type, or ESDT, (EOSDIS Glossary), and the dataset's project, instrument, and mission defined by GCMD Keywords. ESDT is widely used at NASA data archives to uniquely identify datasets and thus is used by many researchers in publications to refer to these datasets. In addition, we add the “NASA” keyword to further eliminate possible ambiguities. The query is thus assembled as “<Dataset ESDT> (<Dataset Platform> | <Dataset Instrument> | <Dataset Project>) NASA”. When searching for the dataset OMBRO (Chance, 2007), the Google Scholar query is: `"/OMBRO/" (/"Aura/" | /"OMI/" ) NASA`, where "Aura" is the satellite platform carrying the "OMI" instrument that collected the data that was used to produce the "OMBRO" dataset.



**Figure 4.** Yearly counts of document citations found by dataset DOI and Keyword search.

Figure 4 presents the document citation counts obtained from bibliometric sources searched by dataset DOIs, and by Google Scholar search via a combination of the dataset ESDT and GCMD Keywords. As can be seen in this figure, the DOI search produced just a few results in earlier years, as datasets had not yet been widely assigned DOIs. In these years the documents are still found by the Google Scholar keywords search. As the number of citations found by dataset DOI search grows over the years, so does the number of citations found by the dataset keywords. This growth tendency shows that the quantity of documents citing datasets by DOI is approaching the number of documents citing datasets by their ESDT name.

## Overlap of Document Citations found by Dataset DOI and Keyword Search



**Figure 5.** Yearly counts of document citations found exclusively by dataset DOI and Keyword search and document citations found by both DOI and Keywords.

The overlap of documents citing datasets by both DOI and mentioning them by ESDT is examined as shown in Figure 5. As seen in Figure 5, the overlap exists, but the quantities of the documents citing datasets by only DOI and only ESDT exceed their overlap. This can be explained by ESDT not being included in the dataset citation, as shown in the current citation for the “OMBRO” dataset:

*Chance, K. (2007), OMI/Aura Bromine Monoxide (BrO) Total Column 1-orbit L2 Swath 13x24 km V003, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), 10.5067/Aura/OMI/DATA2006*

In citations prior to the DOI's assignment to this dataset, the "OMBRO" dataset ESDT name was used:

*Suleiman, R. M., Chance, K., Liu, X., González Abad, G., Kurosu, T. P., Hendrick, F., & Theys, N. (2018). OMI total bromine monoxide (OMBRO) data product: Algorithm, retrieval and measurement comparisons. Atmospheric Measurement Techniques Discuss. <https://doi.org/10.5194/amt-2018-1>*

### **Validation of Google Scholar Results**

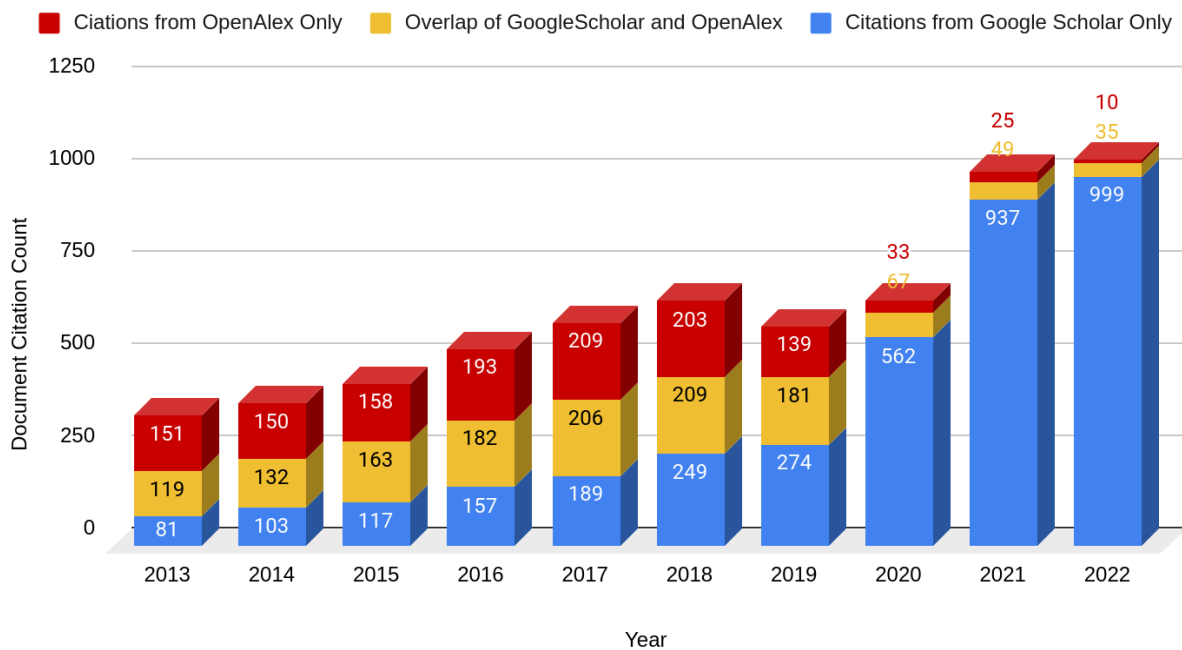
When calculating precision, we considered all articles published in 2021, totaling 945 papers. Out of these, 847 papers, or roughly 90%, had their PDF files available. These PDFs were then converted to ASCII format using the Linux “pdftotext” utility. Each ASCII file was rigorously checked for Short Names, a process guided by findings from Google Scholar. The search within the ASCII files included both specific terms and their variations. Notably, 20 ASCII files did not yield any results from the automated term search, prompting a manual review. Our analysis revealed that Google Scholar's matching capability was superior. For instance, in the paper titled "Aerosol-Cloud Interaction with Summer Precipitation over Major Cities in Eritrea," Google Scholar adeptly detected "TRMM\_3B42\_Daily," even though it was referenced in the paper as "(TRMM) 3B42 (daily)." Furthermore, a mere three files, albeit relevant to the overarching theme, failed to clearly specify the datasets they were built upon. With these findings, we concluded that the precision of our methodology stands at an impressive 99.6%.

### **Comparing Google Scholar's Search Results with OpenAlex**

Google Scholar keyword results were compared with OpenAlex output by searching the latter with the same keyword combinations. For this, OpenAlex search, like Google Scholar's, was limited to papers in English language documents with document types such as books, book chapters, dissertations, and journal and proceedings articles, excluding pre-prints and peer reviews. Only documents that have DOIs were counted. The difference in searches was that for the OpenAlex search the keyword ‘NASA’ was not added. Instead, we relied on the OpenAlex ‘relevance’ score and limited the results to ones having a score higher than 0.99. Relying on the relevance score rather than additional keywords allowed us to produce more hits that were evaluated for relevance by OpenAlex. Since Google Scholar does not

have a relevance score, omitting ‘NASA’ or similar keywords indicating that the data were acquired from the archive produces erroneous results for some of the datasets. As we consider precision as a factor more important than recall, our search emphasis is on fewer results with better precision. Similar to Google Scholar keyword search results, the precision of OpenAlex results was evaluated by examining the documents published in 2018. In 2018 there were 412 documents, for which 381 PDF files were found. Similar to Google Scholar evaluation, PDF files were converted to ASCII files, which in turn were searched for the dataset name keywords. All 381 files contained the keywords, which allowed us to conclude 100% accuracy of the OpenAlex keyword search.

### Overlap of Document Citations Found by Google Scholar and OpenAlex



**Figure 6.** Yearly counts of document citations found exclusively by Google Scholar and OpenAlex

Keyword search and document citations found by both Google Scholar and OpenAlex

The results reported in Figure 6 show that up to the year 2018, OpenAlex slightly outperformed Google Scholar, with its performance significantly degrading after the year 2019.

These results show that as OpenAlex continues improving it should be reevaluated in the future as an alternative solution to Google Scholar. We also looked at using OpenAlex for the referenced dataset

DOI search and as it produced insignificant results, we concluded that it is not yet a viable alternative to Google Scholar.

### **Automated Citation Collection Applications**

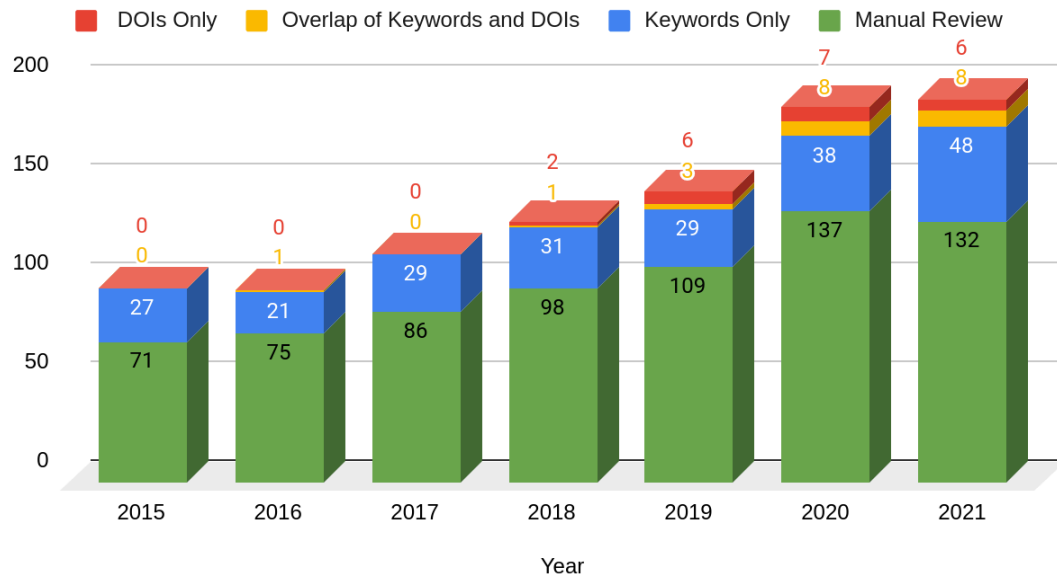
Data providers and science teams that archive datasets at the data center often collect citations of publications about their data collection algorithms, data validation, and overview, as well as publications about applications that use their datasets. A science team's web interface commonly provides access to these citations. It is possible that some of these data citations aren't directly relevant to the datasets because they may be related to general data science, instrumentation, or data from similar projects.

Publications that refer to data collections may also acknowledge the services that were used to acquire those data, such as subsetting, reformatting, visualizing, or analyzing the data collections. NASA Giovanni (Acker and Leptoukh, 2007), developed by GES DISC, has been a popular service for data analysis and visualization of NASA Earth datasets since early 2000. GES DISC regularly collects research publications that mention the usage of the NASA Giovanni service by exhaustive keyword searches of Google Scholar. By the beginning of 2023, around 3,000 research publications have been collected. As GES DISC collects the papers that cite the Giovanni service, the publications are reviewed to determine from which data collections the variables were derived so proper attribution can be assigned to the science teams who created these datasets, and to also list these publications on corresponding dataset web pages at the GES DISC user website.

Using our proposed automated methods, we analyzed how many publications can be found and automatically attributed to datasets, compared to how many publications must be collected using general keywords and manually reviewed to make such attribution. Figure 6 shows counts of citations that were found automatically by DOI and keyword search, their overlap, and counts of publications that were found by generic search and which required manual review.



### Document Citations Using Giovanni Service and Linked to Individual Datasets

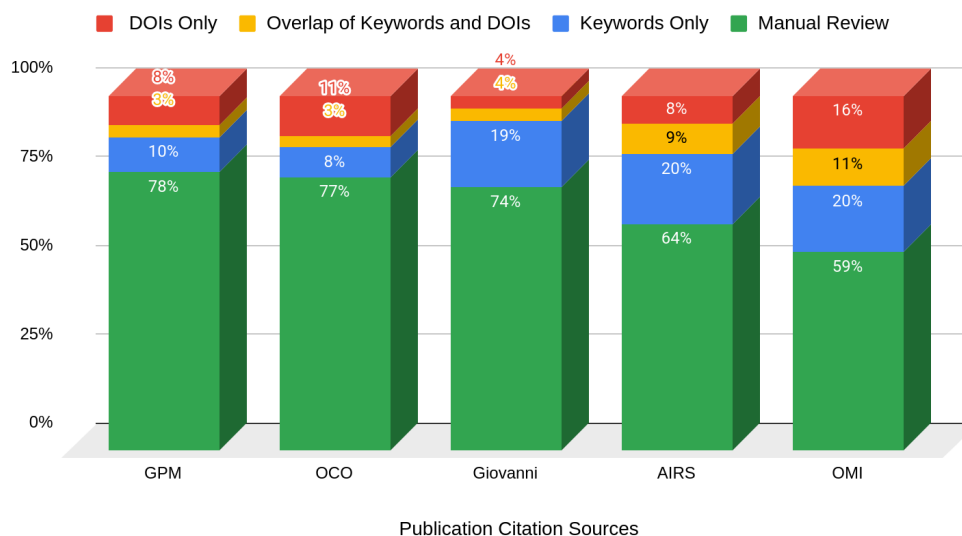


**Figure 7.** Citation counts for dataset attribution to documents using the Giovanni service

Figure 7 demonstrates that a significant count of publications that use GES DISC data still cannot be automatically harvested from bibliometric sources because these documents do not refer to the datasets by DOIs or ESDT names, while still mentioning the NASA Giovanni service used to analyze data. Nevertheless, we can see that at least a quarter of the publications that use data from the Giovanni service can automatically be attributed to the individual datasets.

By comparing data produced by the Atmospheric Infrared Sounder (AIRS), OMI, the Global Precipitation Measurement (GPM) mission, and the Orbiting Carbon Observatory (OCO) with Giovanni's publications, Figure 8 shows how the proposed method can be used to identify datasets in publications collected by several GES DISC data provider teams. The data in Figure 8 covers statistics for research documents published in 2021. As is seen in Figure 8, the proportion of the documents where datasets can be identified with automated search is different. It may be influenced by several factors, one of them being how long the instrument has been operating and providing the data: older missions (AIRS since 2002 and OMI since 2004) have more datasets identified than newer missions (both GPM and OCO since 2014).

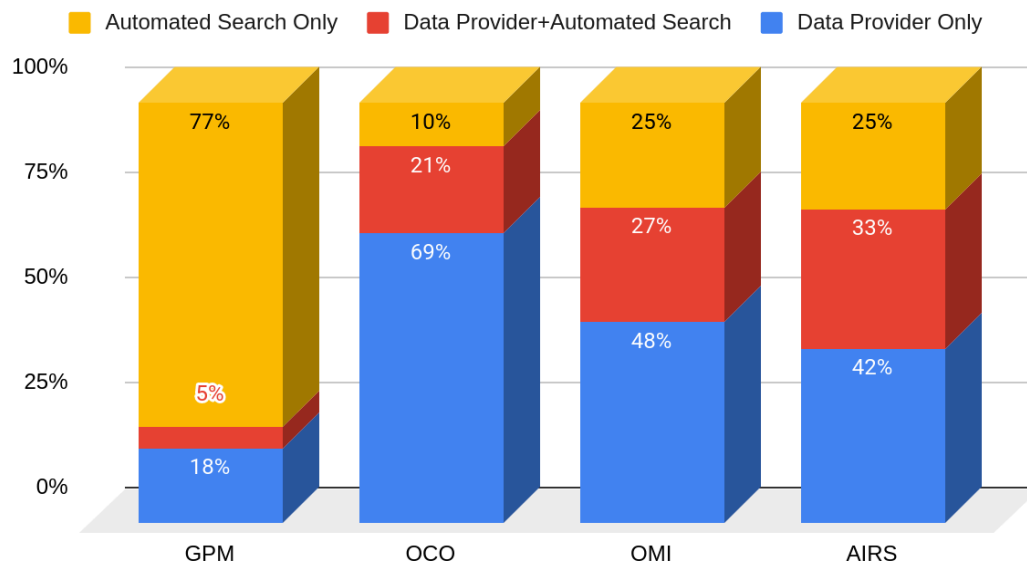
Dataset Attribution to Publications from Data Providers and Giovanni Service



**Figure 8.** Fraction of data provider and Giovanni service documents that can be automatically found and attributed to datasets. The publication year for all these documents is 2021.

Automated search makes it possible to find publications in addition to the publications collected by data providers since automated search finds document citations independently from bibliographic sources. Figure 9 presents the proportion of the document citations that are found by the automated search compared to the total documents collected by the data providers and their search. This proportion varies significantly between the data providers and can be explained by both the data provider publication search methods and resources spent on publication search and review. As publication collection can take significant time to search for and review publications, our method — while it does not find the majority of the publications — can still be used as an automated method to find a relatively constant fraction of publications reliably linked to the datasets.

### Automated Search Citations Add On to Citations Found by Data Provider

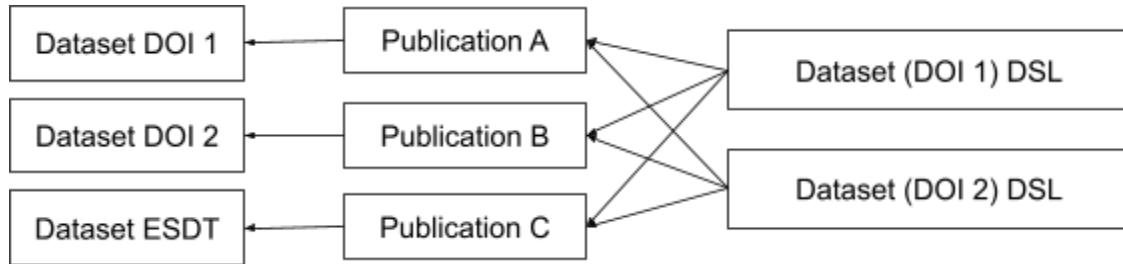


**Figure 9.** Fraction of document citations that are found by an automated search in addition to data provider-found citations. The publication year of all these documents is 2021

### GES DISC Document Citation Management

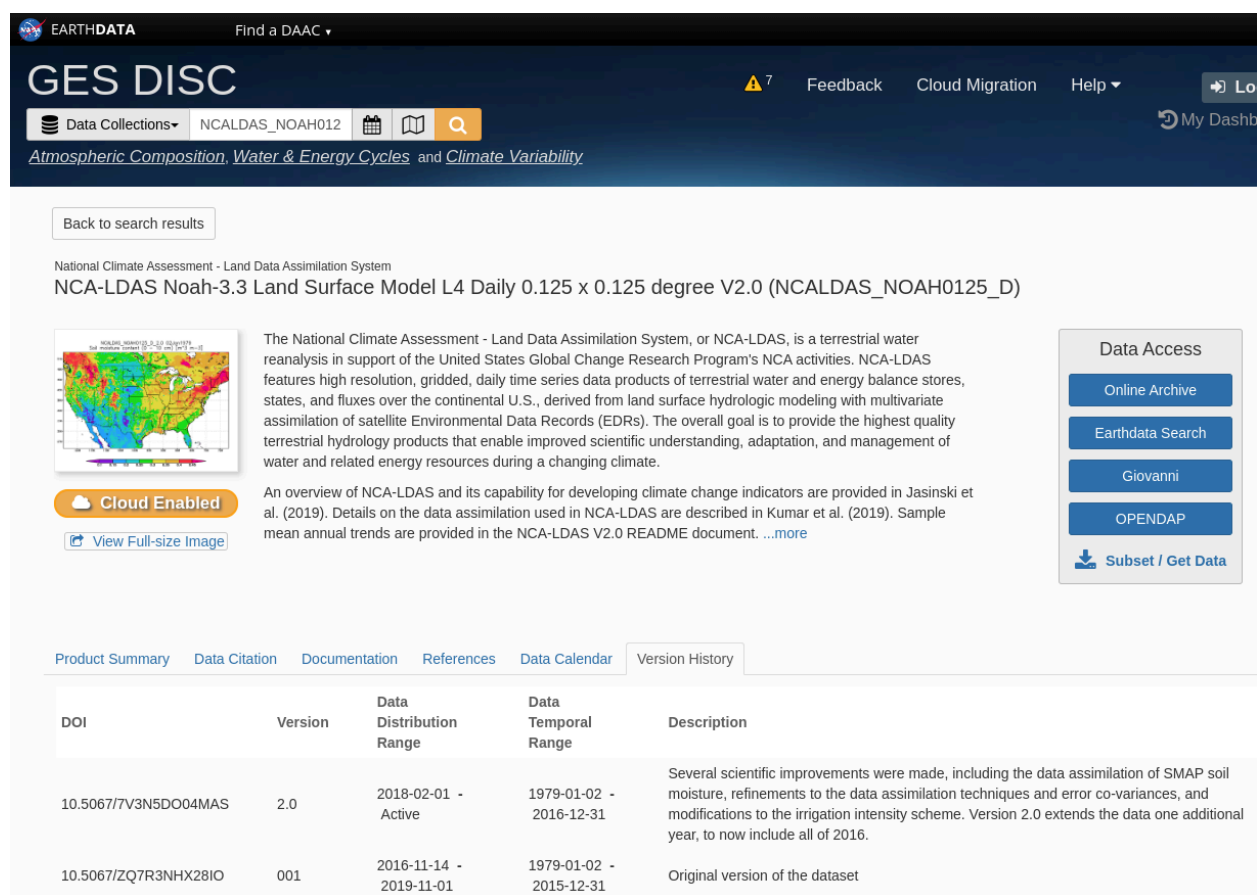
The GES DISC goal is to present collected document citations to archive users in such a way that users would have direct access to all document citations that cite a particular dataset, and the users will have direct access to the datasets when they are presented with the document citation. Several factors complicate this linkage. First, the dataset DOIs are assigned to specific dataset versions. When a dataset is mentioned in the document by its ESDT, it is not always known which version the document authors have used — thus the document cannot be linked to a specific dataset DOI with absolute certainty. Second, older versions of the datasets are retired from the archive, and the DOIs of retired dataset versions are redirected to point to the DOIs of the versions that are distributed to the public. Thus, the publications which used retired versions need to be made available to the users. Third, more than one version of the dataset can be distributed to the users by the archive at the same time, and a newer dataset version usually would have much fewer citations, especially at the beginning of its availability. Thus, if document

citations are linked to the specific version of the datasets, users may miss out on the research literature that cited the prior dataset versions. To address these issues, the GES DISC approach is to link all found documents to the dataset ESDT and present users with all of these documents for each version of the dataset on that dataset version's dataset landing page (DSL), where DSL is the webpage to which the dataset DOI resolves. This approach is presented in Figure 10.



**Figure 10.** Linking publications to datasets at GES DISC

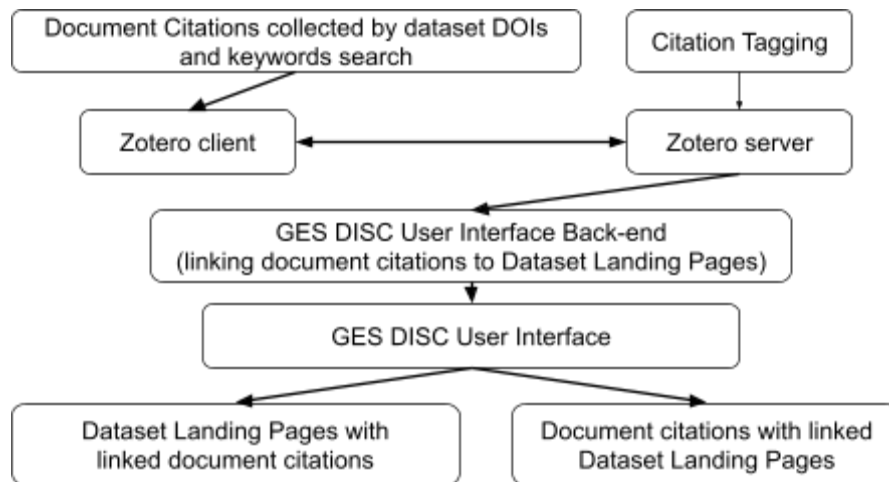
This GES DISC approach to citation linkage lets users see a full history of research literature that used the dataset, regardless of its version. To track previous versions of the dataset, GES DISC implemented a DOI version history that lists all previous DOIs for deprecated dataset versions on the DSL as shown in Figure 11. Documents can be matched with the exact dataset versions used in them when these dataset versions were cited in documents using this DOI history. The version history can help narrow down the versions of datasets that might have been used in documents that didn't cite exact dataset versions.



**Figure 11.** Snapshot of the dataset version history in Dataset Landing Page for the “NCALDAS\_NOAH0125\_D” dataset (Jasinski, 2018)

Dataset registry DataCite links dataset DOIs only to research documents that cited those DOIs. The DataCite has no infrastructure to link datasets to the literature that cited datasets for versions that did not have DOIs, or when the datasets were mentioned and the version that was used was uncertain.

As datasets progress through subsequent versions, the linkage of the documents to the public dataset versions has to be managed. Once a version is retired, the documents are unlinked from it, and once the new version becomes available, all documents linked to that dataset are linked to that version.



**Figure 12.** Flow diagram of the GES DISC document citation management system

Dataset version management related to document linkage is managed separately from document citation collection. Figure 12 presents the diagram of GES DISC document citation management. The document citations are collected from various sources and by different means, and ingested into the Zotero citation management system. Citations are added to Zotero using the Zotero desktop client by import in either bibliographic formats such as BibTeX or by the citation’s DOI. In addition, the Zotero Web browser plugin, called ‘Zotero connector’, is used to load citations directly from the Web browser into the Zotero client.

The Zotero client is connected to the Zotero server, which provides multiple capabilities; it allows for unlimited document citation storage along with citation tagging. The citation library is backed up at the Zotero server and can be shared between individual clients. The library content on the server can be accessed through its API. GES DISC uses the Python API client library, *PyZotero*, to manage the library content for export and citation tagging. GES DISC uses Zotero tags to add the following information to document citations: dataset ESDT name, dataset DOI when cited, and the bibliometric sources where document citation was retrieved.



“Data Collection References” and “Related Publications”. The former are the citations of the publications that the dataset provider advises the GES DISC to supply. These publications typically describe the measurement instrument, dataset algorithm, and validation. “Related Publications” are the publications that are collected by GES DISC that use the dataset for research or investigative purposes. The disclaimer provided for these citations states *“The majority of publications using GES DISC data do NOT include the data version id. As such, publications below may not use the most recent processing version of the data.”* Under each publication there are URLs to DSL pages of linked datasets.

At the GES DISC website, the document citations are also provided at the dedicated search interface as shown in Figure 14. As on the DSL pages under each document citation, there are URLs of corresponding dataset DSLs. The document citations on this interface can be refined by the document publication year, type, and journal. Citations also can be sorted by publication year and author names, as well as downloaded in BibTex format that allows for easy export to the user’s own citation management system. The additional capability of this interface is the search of the publications by the dataset and free text keywords. Free text search on the publications interface opens the possibility of data discovery by their application. For example, if a user searches for “landslide”, then the search results will list documents that have this keyword appear in their title and/or abstract and which very likely relate to the research related to landslides. Direct linkage of these citations to the datasets allows the user to “discover” datasets that were used for research on landslides.



**EARTHDATA** Find a DAAC

# GES DISC

Publications Enter search (e.g., rainfall) Feedback Cloud Migration Help

Atmospheric Composition, Water & Energy Cycles and Climate Variability

**Publications** Showing 1 - 25 of 9537 publications Download as BibTeX Sort by: Year

**Refine By**

**Article Type** Sort

- ☐ article-journal (8916)
- ☐ book (229)
- ☐ conference (308)
- ☐ report (50)
- ☐ thesis (33)
- More...

**Journal** Sort

- ☐ ACS Earth and Space Chemistry (4)
- ☐ ACS Sustainable Chemistry & Engineering (1)
- ☐ ADVANCES IN WATER RESOURCES (1)
- ☐ AEROSOL AND AIR QUALITY RESEARCH (2)
- ☐ AGRICULTURAL AND FOREST METEOROLOGY (2)
- More...

**Year** Sort

- ☐ 2023 (89)
- ☐ 2022 (1161)
- ☐ 1997 (8)
- ☐ 1996 (12)

**2023**

Piñango, Andres, Azar, Elias, Wallner-Kersanach, Mônica, da Costa Machado, Eunice, Martins, Gabriela, Peterle, Thayná, de Rezende, Carlos Eduardo, da Graça Baumgarten, Maria. 2023. **Influence of the ITCZ and OMZ on the isotopic composition of suspended particulate matter in the western tropical North Atlantic.** *Journal of Marine Systems*. Vol. 237, pp. 103803. DOI: [10.1016/j.jmarsys.2022.103803](https://doi.org/10.1016/j.jmarsys.2022.103803) ISSN: 09247963

**Related Data Collections (3)**

- [GPCPMON\\_3.1](#)
- [GPCPMON\\_3.0](#)
- [GPCPMON\\_3.2](#)

Saha, Prolay, Mahanta, Rahul, Rajesh, P. V., Goswami, B. N.. 2023. **Persistent Wet and Dry Spells of Indian Summer Monsoon Rainfall: A Reexamination of Definitions of "Active" and "Break" Events.** *Journal of Climate*. Vol. 36, No. 1, pp. 261-277. DOI: [10.1175/JCLI-D-21-1003.1](https://doi.org/10.1175/JCLI-D-21-1003.1) ISSN: 0894-8755, 1520-0442

**Related Data Collections (1)**

- [TRMM\\_3B42\\_Daily\\_7](#)

Setti, Sridhara, Yumnam, Karisma, Rathinasamy, Maheswaran, Agarwal, Ankit. 2023. **Assessment of satellite precipitation products at different time scales over a cyclone prone coastal river basin in India.** *Journal of Water and Climate Change*. Vol. 14, No. 1, pp. 38-65. DOI: [10.2166/wcc.2022.166](https://doi.org/10.2166/wcc.2022.166) ISSN: 2040-2244, 2408-9354

**Related Data Collections (1)**

- [TRMM\\_3B42\\_Daily\\_7](#)

**Figure 14.** Snapshot of GES DISC Publications interface

<https://disc.gsfc.nasa.gov/information/publications>

## Summary

We developed an automated approach to obtain and process large quantities of citations from Google Scholar and to accurately link them with the datasets used in the cited research. To develop this approach, we first analyzed all major bibliographical resources such as Web of Science, Scopus, Crossref, and Google Scholar. We concluded that the latter not only covers the majority of citations that can be found in other sources but also contains citations missing from the rest. Based on this knowledge, we then developed a method of obtaining document citations from Google Scholar using the subscription-based SerpAPI. The major components of our method consist of composing dataset-specific keyword search queries, and subsequent application of open-source Zotero translation software, for converting obtained

URLs into proper document citations. These valid citations are then further processed to deduplicate them and retain citations with desired document types. We evaluated the precision of our method for all citations published in 2021 by verifying that those papers indeed referenced or mentioned the datasets. To showcase the comparative efficacy of our method against other available solutions, we implemented a dataset keyword search using OpenAlex. Our findings indicate that OpenAlex stands as a promising tool, adept at facilitating efficient text searches within publications. We illustrated the capacity of our method to enhance the efficiency of collection citations, utilizing a control sample gathered from dataset provider science teams as a benchmark. Our results show that for some of the data provider citation websites, as many as 40% of their citations can be found by our automated approach. Finally, we demonstrate the workflow of how the document-dataset linked citations are collected at the GES DISC and how potential data users benefit from this process.

## **Acknowledgments**

This work was funded under the NASA SESDA IV Contract No. 80GSFC17C0003.

## **Computer Code Availability**

The Google Scholar Dataset Citing Documents Search (GSDCDS) tool is available from GitHub. The document citations collected for the GES DISC archive are maintained in the GES DISC Zotero library.

## **References**

Acker, J. G., Leptoukh, G. (2007). Online analysis enhances use of NASA Earth science data. Eos, Transactions American Geophysical Union, 88(2), 14–17. <https://doi.org/10.1029/2007EO020003>

AIRS - Atmospheric Infrared Sounder - Publications, (n. d.). Retrieved April 20, 2023, from <https://airs.jpl.nasa.gov/resources/publications>

- Behnke, J., Mitchell, A., & Ramapriyan, H. (2019). NASA's Earth Observing Data and Information System – Near-Term Challenges. *Data Science Journal*, 18(1), Article 1.  
<https://doi.org/10.5334/dsj-2019-040>
- Böhner, D., Teichert, B. (2020). Reference Management Software Comparison - 8th update, Technical University of Munich, University Library, permanent link: <http://mediatum.ub.tum.de/1320978>
- Brase, J. (2009). DataCite - A Global Registration Agency for Research Data, *Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology*, Beijing, 2009, pp. 257-261, <https://doi.org/10.1109/COINFO.2009.66>
- Burnham, J. F. (2006). Scopus database: A review. *Biomedical Digital Libraries*, 3(1), 1.  
<https://doi.org/10.1186/1742-5581-3-1>
- Chance, K. (2007), OMI/Aura Bromine Monoxide (BrO) Total Column 1-orbit L2 Swath 13x24 km V003, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), <https://doi.org/10.5067/Aura/OMI/DATA2006>
- Chapman, K., & Ellinger, A. E. (2019). An evaluation of Web of Science, Scopus, and Google Scholar citations in operations management. *The International Journal of Logistics Management*, 30(4), 1039–1053. <https://doi.org/10.1108/IJLM-04-2019-0110>
- CMR - Common Metadata Repository, (n. d.) Retrieved May 2023  
<https://www.earthdata.nasa.gov/eosdis/science-system-description/eosdis-components/cmr>
- CODATA-ICSTI - Task Group on Data Citation Standards and Practices. (2013). Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. *Data Science Journal*, 12, pp. 75, <https://doi.org/10.2481/dsj.OSOM13-043>
- Costas, R., Meijer, I., Zahedi, Z., & Wouters, P. F. (2013). *The value of research data metrics for datasets from a cultural and technical point of view. A knowledge exchange report*. Leiden. Retrieved June 6, 2022, from <https://hdl.handle.net/1887/23586>

Cousijn, H., Feeney, P., Lowenberg, D., Presani, E., & Simons, N. (2019). Bringing Citations and Usage Metrics Together to Make Data Count. *Data Science Journal*, 18(1), 9.

<https://doi.org/10.5334/dsj-2019-009>

Delgado López-Cózar, E., Orduna-Malea, E., & Martín-Martín, A. (2018). Google Scholar as a data source for research assessment (arXiv:1806.04435). arXiv.

<https://doi.org/10.48550/arXiv.1806.04435>

Duan, X., Zhang, J., Ramachandran, R., Gatlin, P., Maskey, M., Miller, J. J., Bugbee, K., & Lee, T. J. (2018). A Neural Network-Powered Cognitive Method of Identifying Semantic Entities in Earth Science Papers. *2018 IEEE International Conference on Cognitive Computing (ICCC)*, 9–16,

<https://doi.org/10.1109/ICCC.2018.00009>.

EOSDIS Glossary, (n. d.) Retrieved April 30, 2023, <https://www.earthdata.nasa.gov/learn/glossary>

GSDCDS - Google Scholar Dataset Citing Documents Search tool, v1.0 (n. d.),

<https://github.com/iragerasimov/GSDCDS>

GCMD - Global Change Master Directory Keywords (n. d.) Retrieved May 9, 2023,

<https://www.earthdata.nasa.gov/learn/find-data/idn/gcmd-keywords>

GES DISC, (n. d.) Retrieved April 30, 2023, <https://disc.gsfc.nasa.gov/>.

GES DISC Zotero library, (n. d.) Retrieved September 10, 2023, at

[https://www.zotero.org/groups/2395775/ges\\_disc/library](https://www.zotero.org/groups/2395775/ges_disc/library).

GPM - Global Precipitation Measurement - Publications, (n. d.). Retrieved April 20, 2023, from

<https://pmm.nasa.gov/resources/gpm-publications>

Gusenbauer, M. (2019). Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, 118(1), 177–214.

<https://doi.org/10.1007/s11192-018-2958-5>

I4OA - Initiative for Open Abstracts, (n. d.) Retrieved April 30, 2023, <https://i4oa.org/>

I4OC - Initiative for Open Citations, (n. d.) Retrieved April 30, 2023, <https://i4oc.org/>

- Jasinski, M.F., S.V. Kumar, J.S. Borak, D.M. Mocko, C.D. Peters-Lidard, M. Rodell, H. Rui, H. Kato Beaudoin, B.E. Vollmer, K.R. Arsenault, B. Li, and J.D. Bolten (2018), NCA-LDAS Noah-3.3 Land Surface Model L4 Daily 0.125 x 0.125 degree V2.0, Greenbelt, Maryland, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC),  
<https://doi.org/10.5067/7V3N5DO04MAS>
- Halevi, G., Moed, H., & Bar-Ilan, J. (2017). Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation - Review of the Literature. *Journal of Informetrics*, 11(3), 823–834. <https://doi.org/10.1016/j.joi.2017.06.005>
- Heibi, I., Peroni, S., & Shotton, D. (2019). Software review: COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations. *Scientometrics*, 121(2), 1213–1228.  
<https://doi.org/10.1007/s11192-019-03217-6>
- Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, 1(1), 414–427.  
[https://doi.org/10.1162/qss\\_a\\_00022](https://doi.org/10.1162/qss_a_00022)
- Ivey, C., & Crum, J. (2018). Choosing the Right Citation Management Tool: Endnote, Mendeley, Refworks, or Zotero. *Journal of the Medical Library Association : JMLA*, 106(3), 399–403,  
<https://doi.org/10.5195/jmla.2018.468>.
- Kratz, J. E., & Strasser, C. (2015). Researcher Perspectives on Publication and Peer Review of Data. *PLOS ONE*, 10(2), 21. <https://doi.org/10.1371/journal.pone.0117619>
- Lane, J., Mulvany, I., Nathan, P. (2020). *Rich Search and Discovery for Research Datasets: Building the Next Generation of Scholarly Infrastructure*. London: Sage, ISBN 978-1-5297-0586-7.
- Li, X., Thelwall, M. (2012). F1000, Mendeley and traditional bibliometric indicators. *Proceedings of the 17th International Conference on Science and Technology Indicators*, 2, 451–551.
- Martín-Martín, A., Orduna-Malea, E., Thelwall, M., Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, 12(4), 1160–1177, doi: 10.1016/j.joi.2018.09.002.

Mooney, H., & Newton, M. P. (2012). The Anatomy of a Data Citation: Discovery, Reuse, and Credit.

*Journal of Librarianship and Scholarly Communication*, 1(1). <https://doi.org/20210729115526>

OCO - Orbiting Carbon Observatory - Publications, (n. d.). Retrieved April 20, 2023, from

<https://ocov2.jpl.nasa.gov/science/publications/>

OMI Publications, (n. d.). Retrieved April 20, 2023, from

<https://acd-ext.gsfc.nasa.gov/Documents/Publications/OMI/>

Park, H., & Wolfram, D. (2017). An examination of research data sharing and re-use: Implications for data citation practice. *Scientometrics*, 111(1), 443–461.

<https://doi.org/10.1007/s11192-017-2240-2>

Parsons, M. A., Duerr, R. E., & Jones, M. B. (2019). The History and Future of Data Citation in Practice.

*Data Science Journal*, 18(1), <https://doi.org/10.5334/dsj-2019-052>

Paskin, N. (2002). Digital object identifiers. *Information Services & Use*, 22(2), 97–112,

<https://doi.org/10.3233/ISU-2002-222-309>.

Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. (2016). Research data explored: An extended analysis of citations and altmetrics. *Scientometrics*, 107(2), 723–744.

<https://doi.org/10.1007/s11192-016-1887-4>

Prins, A. A. M., Costas, R., van Leeuwen, T. N., & Wouters, P. F. (2016). Using Google Scholar in research evaluation of humanities and social science programs: A comparison with Web of Science data. *Research Evaluation*, 25(3), 264–270. <https://doi.org/10.1093/reseval/rvv049>

Robinson-García, N., Jiménez-Contreras, E., Torres-Salinas, D. (2016). Analyzing data citation practices using the data citation index. *Journal of the Association for Information Science and Technology*,

67(12), 2964–2975. <https://doi.org/10.1002/asi.23529>

Robinson-Garcia, N., Mongeon, P., Jeng, W., & Costas, R. (2017). DataCite as a novel bibliometric source: Coverage, strengths and limitations. *Journal of Informetrics*, 11(3), 841–854,

<https://doi.org/10.1016/j.joi.2017.07.003>.

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts (arXiv:2205.01833). arXiv.

<https://doi.org/10.48550/arXiv.2205.01833>

Roy Rosenberg Center for History and New Media, George Mason University (2020). Zotero. Available:

<https://www.zotero.org/>.

Silvello, G. (2018). Theory and practice of data citation. *Journal of the Association for Information Science and Technology*, 69(1), 6–20. <https://doi.org/10.1002/asi.23917>

Van Noorden, R. (2014). Google Scholar pioneer on search engine's future. *Nature*.

<https://doi.org/10.1038/nature.2014.16269>

Vannan, S., Downs, R. R., Meier, W., Wilson, B. E., & Gerasimov, I. V. (2020), Data sets are foundational to research. Why don't we cite them?, *Eos*, 101, <https://doi.org/10.1029/2020EO151665>.

Wanchoo, L., James, N., & Ramapriyan, H. (2017). NASA EOSDIS Data Identifiers: Approach and System. *Data Science Journal*, 16(0), Article 0. <https://doi.org/10.5334/dsj-2017-015>

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Santos, L. B. da S., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 1–9, <https://doi.org/10.1038/sdata.2016.18>.

Zhao, M., Yan, E., & Li, K. (2018). Data set mentions and citations: A content analysis of full-text publications. *Journal of the Association for Information Science and Technology*, 69(1), 32–46.

<https://doi.org/10.1002/asi.23919>